D1-82-0152

BOEING SCIENTIFIC RESEARCH LABORATORIES

# Uniform Distributions Over a Simplex

G. Marsaglia

Mathematics Research



December, 1961

D1-82-0152

# UNIFORM DISTRIBUTIONS OVER A SIMPLEX

by

G. Marsaglia

Mathematical Note No. 250

Mathematics Research Laboratory

BOEING SCIENTIFIC RESEARCH LABORATORIES

December 1961

# UNIFORM DISTRIBUTIONS OVER A SIMPLEX

## 1. Introduction

We are interested in a random assignment of $n + 1$ probabilities $p_1,\ldots,p_{n+1}$, with $p_1 + \cdots + p_{n+1} = 1$. By that we mean that the random vector $\pi = (p_1,\ldots,p_{n+1})$ is uniformly distributed over the simplex $S_{n+1} = \{(x_1,\ldots,x_{n+1}): x_i \geq 0, \sum_1^{n+1} x_i = 1\}$, and this, in turn, has a meaning we will make precise in section 2. We will find the joint distribution of the $p$'s and consider various ways of representing the $p$'s in terms of familiar random variables. It turns out that the joint distribution of the $p$'s is very simple when given in terms of $P[p_1 > a_1,\ldots,p_{n+1} > a_{n+1}]$ rather than the conventional $P[p_1 < a_1,\ldots,p_{n+1} < a_{n+1}]$, in fact, $P[p_1 > a_1,\ldots,p_{n+1} > a_{n+1}] = (1 - a_1 - \cdots - a_{n+1})^n$.

The principal purpose of the development is to get Theorem 4, which provides a method for generating exponential (and hence normal, using polar coordinates) random variables in a computer -- $\pi$ is chosen from $S_{n+1}$, then a random variable $z$ is chosen so that $zp_1, zp_2,\ldots,zp_{n+1}$ are independent and exponentially distributed. We will also point out how $\pi$ may be used to produce points uniformly over the simplex supported by an arbitrary set of $n + 1$ linearly independent points in space. Thus the main purpose of the development is to provide some theory on which certain "Monte Carlo" techniques may be based, but the development will also provide an approach to problems involving uniform order statistics which is different from that usually taken.

## 2. Uniform Distributions

Let $R_n$ be Euclidean n-space, and let $\lambda$ be Lebesgue measure. We say that $z_1,\ldots,z_n$ are jointly uniformly distributed, or that the vector $\zeta = (z_1,\ldots,z_n)$ is uniformly distributed, over a subset $S \subset R_n$ with positive Lebesgue measure, if the probability measure for $\zeta$ is proportional to Lebesgue measure -- for $\lambda$-measurable $B$, $P[(z_1,\ldots,z_n)\varepsilon B] = \lambda(B)/\lambda(S)$. Then $z_1,\ldots,z_n$ have a joint density function; it is constant, $1/\lambda(S)$, over $S$ and zero elsewhere. If $A$ is the $n \times n$ non-singular matrix of a linear transformation $T$, then $T(\zeta)$, $(=\zeta A)$, is uniformly distributed over the image set $T(S)$.

Let $G$ be the simplex supported by $k$ linearly independent points $\alpha_1,\ldots,\alpha_k$ in $R_k$: $G = \{\beta: \beta = \sum_1^k c_i\alpha_i,\ c_i \geq 0,\ \sum_1^k c_i = 1\}$. We say that $\eta = (y_1,\ldots,y_k)$ is uniformly distributed over the simplex $G$ if for some $i$ the vector $(y_1,y_2,\ldots,y_{i-1},y_{i+1},\ldots,y_k)$ is uniformly distributed, in the above sense, over the projection of $G$ onto $R_{k-1}$ formed by suppressing the ith coordinate, i.e., $T_i(G)$, where $T_i$ is the transformation $(y_1,\ldots,y_k) \longrightarrow (y_1,y_2,\ldots,y_{i-1},y_{i+1},\ldots,y_k)$ taking $R_k$ onto $R_{k-1}$. This is equivalent to saying that the probability that $\eta$ fall in a subset $H$ of $G$ is proportional to the surface area of $H$. For each $i$, $T_i(\eta)$ will be uniformly distributed in the first sense over $T_i(G)$, except in the nuisance case when one of the coordinates, say $x_j$, of each point $\xi = (x_1,\ldots,x_k)$ of $G$ is constant; $x_j = c$. Then only $T_j(\eta)$ will be uniformly distributed.

We note in particular that if $\pi = (p_1,\ldots,p_{n+1})$ is uniformly distributed over

$$S_{n+1} = \{(x_1,\ldots,x_{n+1}): x_i \geq 0, \sum_1^{n+1} x_i = 1\},$$

then each set of $n$ of the $p$'s is uniformly distributed over the regular polytope

$$S_n^* = \{(x_1,\ldots,x_n): x_i \geq 0, \sum_1^n x_i \leq 1\}.$$

If $A$ is the $k \times k$ non-singular matrix of a linear transformation $T$, then $T(\eta)$, $(=\eta A)$, is uniformly distributed over the simplex $T(G)$. This follows from noting that $T$ is the product of elementary linear transformations, and if $y_1,\ldots,y_k$ are uniformly distributed over a simplex, then $ay_1,y_2,\ldots,y_k$ $(a \neq 0)$, or $y_2,y_1,y_3,\ldots,y_k$ or $y_1 + cy_2,y_2,\ldots,y_k$ are uniformly distributed over the appropriately transformed simplex. We have in particular:

Theorem 1. If $\pi = (p_1,\ldots,p_{n+1})$ is uniformly distributed over the simplex $S_{n+1} = \{(x_1,\ldots,x_{n+1}): x_i \geq 0, \sum_1^{n+1} x_i = 1\}$ and if $\alpha_1,\ldots,\alpha_{n+1}$ are $n+1$ linearly independent points in $R_{n+1}$, then the random vector $p_1\alpha_1 + p_2\alpha_2 + \cdots + p_{n+1}\alpha_{n+1}$ is uniformly distributed over $G$, the simplex supported by the $\alpha$'s: $G = \{\gamma: \gamma = \sum_1^{n+1} c_i\alpha_i, c_i \geq 0, \sum_1^{n+1} c_i = 1\}$.

This theorem provides a practical method for producing random points uniformly over regions with linear boundaries, for example, in certain Monte Carlo procedures, and it also shows how the distributions of the coordinates of a random uniform point from a simplex may be found in terms of the distribution of a linear combination of $p_1,\ldots,p_{n+1}$.

If $\gamma_1,\ldots,\gamma_n$ are $n$ linearly independent points in $R_n$ and $(p_1,\ldots,p_n)$ is uniformly distributed over $S_n^* = \{(x_1,\ldots,x_n): x_i \geq 0, \Sigma x_i \leq 1\}$, then $p_1\gamma_1 + \cdots + p_n\gamma_n$ is uniformly distributed over the set $H = \{\beta: \beta = \Sigma c_i\gamma_i, c_i \geq 0, \Sigma c_i \leq 1\}$. Let $\gamma_0$ be the origin. Then $H$ may be viewed as the convex polytope supported by $\gamma_0,\gamma_1,\ldots,\gamma_n$, i.e.,

$$H = \{\delta: \delta = c_1\gamma_1 + \cdots + c_n\gamma_n + c_{n+1}\gamma_0,\ c_i \geq 0,\ \sum_1^{n+1} c_i = 1\}.$$

Hence, if $\beta_1 - \beta_0,\ldots,\beta_n - \beta_0$ are linearly independent, and if $\pi = (p_1,\ldots,p_n,p_{n+1})$ is uniformly distributed over

$$S_{n+1} = \{(x_1,\ldots,x_{n+1}): x_i \geq 0,\ \sum_1^{n+1} x_i = 1\},$$

then $p_1\beta_1 + \cdots + p_n\beta_n + p_{n+1}\beta_0 = p_1(\beta_1 - \beta_0) + p_2(\beta_2 - \beta_0) + \cdots + p_n(\beta_n - \beta_0) + \beta_0$ is uniformly distributed over the convex hull of $\beta_0,\beta_1,\ldots,\beta_n$, that set being the translate of the convex hull of $\beta_0 - \beta_0,\beta_1 - \beta_0,\ldots,\beta_n - \beta_0$.

Theorem 2. If $\pi = (p_1,\ldots,p_{n+1})$ is uniformly distributed over the simplex $S_{n+1} = \{(x_1,\ldots,x_{n+1}): x_i \geq 0, \Sigma x_i = 1\}$ and if $\beta_0,\beta_1,\ldots,\beta_n$ are $n+1$ points in $R_n$ such that $\beta_1 - \beta_0,\ldots,\beta_n - \beta_0$ are linearly independent, then the random vector $p_1\beta_1 + \cdots + p_n\beta_n + p_{n+1}\beta_0$ is uniformly distributed over $H$, the convex polytope supported by the $\beta$'s:

$$H = \{\delta: \delta = d_1\beta_1 + \cdots + d_n\beta_n + d_{n+1}\beta_0,\ d_i \geq 0,\ \sum_1^{n+1} d_i = 1\}.$$

The most common application of this result is in the plane -- to choose a point $\beta$ from the triangle formed by $\beta_0,\beta_1,\beta_2$ we choose $(p_1,p_2,p_3)$ uniformly from $S_3$ and put $\beta = p_1\beta_1 + p_2\beta_2 + p_3\beta_0$. To choose a point uniformly from a polygonal region in the plane, we divide

the region into triangles, choose a triangle with probability proportional to its area, then choose a point uniformly from that triangle.

If $(p_1,\ldots,p_n)$ is uniformly distributed over the polytope $S_n^* = \{(x_1,\ldots,x_n)\colon x_i \geq 0, \Sigma\, x_i \leq 1\}$, then the vector $\omega = (p_1, p_1 + p_2, \ldots, p_1 + p_2 + \cdots + p_n)$ is uniformly distributed over $Q_n = \{(y_1,\ldots,y_n)\colon 0 \leq y_1 \leq y_2 \leq \cdots \leq y_n \leq 1\}$ and hence the components of $\omega$ are distributed as the order statistics of independent uniform $[0,1]$ random variables $u_1,\ldots,u_n$. Going in the other direction, if $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(n)}$ are the order statistics of $n$ independent uniform $[0,1]$ random variables, then the vector $(u_{(1)}, u_{(2)} - u_{(1)}, \cdots, u_{(n)} - u_{(n-1)})$ is uniformly distributed over $S_n^*$ and the vector $(u_{(1)}, u_{(2)} - u_{(1)}, \cdots, u_{(n)} - u_{(n-1)}, 1 - u_{(n)})$ is uniformly distributed over

$$S_{n+1} = \{(x_1,\ldots,x_{n+1})\colon x_i \geq 0, \sum_1^{n+1} x_i = 1\}.$$

This relationship provides a practical method for producing $\pi$ uniformly on $S_{n+1}$ in a computer -- we order a set of $n$ independent uniform random variables and take successive differences.

We will need the following fact: if $u_1,\ldots,u_n$ are independent random variables, each uniform on $[0,1]$, and $0 \leq c \leq 1$, then

$$P[u_1 + \cdots + u_n < c] = c^n/n! \quad . \tag{1}$$

This is a special case of the well-known result on the distribution of the sum of uniform random variables, but it can be proved by induction

quite easily, since

$$P[u_1+\cdots+u_n<c] = \int_0^1 P[u_1+\cdots+u_n<c\,|\,u_n=x]dx = \int_0^c P[u_1+\cdots+u_{n-1}<1-x]dx.$$

We note that (1) gives the Lebesgue measure of the region

$$(2)\qquad A = \{(x_1,\ldots,x_n): x_1>0,\ldots,x_n>0,\ x_1+\cdots+x_n<c\}.$$

3. The Joint Distribution of $p_1,\ldots,p_{n+1}$

Theorem 3. If $\pi = (p_1,\ldots,p_n,p_{n+1})$ is uniformly distributed over $S_{n+1} = \{(x_1,\ldots,x_{n+1}): x_i>0,\ \sum_1^{n+1} x_i = 1\}$, then the joint distribution of the p's is, for non-negative $a_1,\ldots,a_{n+1}$:

$$P[p_1>a_1,p_2>a_2,\ldots,p_n>a_n,p_{n+1}>a_{n+1}] = (1-a_1-a_2-\cdots-a_{n+1})^n$$

for $a_1+\cdots+a_{n+1}\le 1$ and zero for $a_1+\cdots+a_{n+1}>1$.

The proof follows readily from the fact that the two sets

$$A = \{(x_1,\ldots,x_n): x_1>0,\ldots,x_n>0, x_1+\cdots+x_n<c\}$$

$$B = \{(y_1,\ldots,y_n): y_1-a_1>0,\ldots,y_n-a_n>0,(y_1-a_1)+(y_2-a_2)+\cdots+(y_n-a_n)<c\}$$

are congruent, and, by (2) and the remark preceding (2), their common Lebesgue measure is $c^n/n!$, for $0\le c\le 1$. Then

$$P[p_1>a_1,\ldots,p_{n+1}>a_{n+1}] = P[p_1>a_1,\ldots,p_n>a_n,1-p_1-\cdots-p_n>a_n]$$

$$= P[p_1-a_1>0,\ldots,p_n-a_n>0,(p_1-a_1)+\cdots+(p_n-a_n)<1-a_1-\cdots-a_{n+1}].$$

The latter is the probability that $(p_1,\ldots,p_n)$ will fall in B, with $c = 1-a_1-\cdots-a_n-a_{n+1}$, and hence equals $\lambda(B)/\lambda(S_n^*) = c^n$.

Putting the appropriate a's equal to zero, we have

Corollary. If $\pi = (p_1,\dots,p_n,p_{n+1})$ is uniformly distributed over $S_{n+1}$, then the joint distribution of $p_1,\dots,p_k,(k \le n)$, is given by

$$P[p_1 > a_1,\dots,p_k > a_k] = (1 - a_1 - a_2 - \cdots - a_k)^n,\ a_i > 0, a_1 + \cdots + a_k \le 1,$$

and, taking partial derivatives, the joint density of $p_1,\dots,p_k$ is

$$n(n+1)\cdots(n-k+1)(1-x_1-\cdots-x_k)^{n-k} \quad \text{for} \quad x_i > 0,\ x_1+\cdots+x_k < 1$$

and zero elsewhere.

We now get this interesting result:

Theorem 4. If $\pi = (p_1,\dots,p_{n+1})$ is uniformly distributed over $S_{n+1} = \{(x_1,\dots,x_{n+1}): x_i \ge 0,\ \sum_1^{n+1} x_i = 1\}$, and if $z$ has density function $x^n e^{-x}/n!$, $x > 0$, and is independent of $\pi$, then the random variables

$$zp_1, zp_2, \dots, zp_{n+1}$$

are independent and exponentially distributed, (density $e^{-x}$, $x \ge 0$).

The proof comes from integrating the conditional probability for given $z$:

$$P[zp_1 > b_1,\dots,zp_{n+1} > b_{n+1}] = \int_0^\infty P[zp_1 > b_1,\dots,zp_{n+1} > b_{n+1} | z = x]\, \frac{x^n e^{-x}}{n!}\,dx$$

$$= \int_0^\infty P[p_1 > \frac{b_1}{x},\dots,p_{n+1} > \frac{b_{n+1}}{x}]\, \frac{x^n e^{-x}}{n!}\,dx = \int_{b_1+\cdots+b_{n+1}}^\infty (1 - \frac{b_1}{x} - \cdots - \frac{b_{n+1}}{x})^n\, \frac{x^n e^{-x}}{n!}\,dx.$$

Letting $c = b_1 + \cdots + b_{n+1}$, and $y = x - c$, we have

$$\int_c^\infty \frac{(x-c)^n e^{-x}}{n!}dx = e^{-c}\int_0^\infty \frac{y^n e^{-y}}{n!}dy = e^{-c}.$$

Thus $P[zb_1 > b_1,\ldots,zp_{n+1} > b_{n+1}] = e^{-b_1-\cdots-b_{n+1}}$, which means that $zp_1,\ldots,zp_{n+1}$ are independent and each has density function $e^{-x}$.

Noting that $\frac{zp_i}{zp_1+\cdots+zp_{n+1}} = p_i$, we have a result in the opposite direction:

Corollary. If $y_1,\ldots,y_{n+1}$ are independent, exponential random variables, then

$$\frac{y_1}{y_1+\cdots+y_{n+1}},\ \frac{y_2}{y_1+\cdots+y_{n+1}},\ \ldots,\ \frac{y_{n+1}}{y_1+\cdots+y_{n+1}}$$

are uniformly distributed over the simplex

$$S_{n+1} = \{(x_1,\ldots,x_{n+1}): x_i \geq 0,\ \sum_1^{n+1} x_i = 1\}.$$

The representation of $p_1,\ldots,p_{n+1}$ in terms of independent exponential random variables $y_1,\ldots,y_{n+1}$ provides an elementary method for establishing well-known properties of uniform order statistics. For example, let

$$v_1 = \frac{y_1}{y_1+\cdots+y_{n+1}},\ v_2 = \frac{y_1+y_2}{y_1+\cdots+y_{n+1}},\ \ldots,\ v_n = \frac{y_1+\cdots+y_n}{y_1+\cdots+y_{n+1}}.$$

Then $(v_1,v_2,\ldots,v_n)$ are uniformly distributed over the polytope $Q_n = \{(x_1,\ldots,x_n): 0 < x_1 < x_2 < \cdots < x_n < 1\}$ and hence are distributed

as the order statistics of $n$ independent uniform $[0,1]$ random variables.

Now let $k$ be given, $1 \leq k \leq n$. Then $v_1,\ldots,v_{k-1}$ may be written

$$(5) \qquad \frac{y_1}{y_1+\cdots+y_k}v_k,\ \frac{y_1+y_2}{y_1+\cdots+y_k}v_k,\ \ldots,\ \frac{y_1+\cdots+y_{k-1}}{y_1+\cdots+y_k}v_k,$$

and $v_{k+1},\ldots,v_n$ may be written

$$(6) \qquad v_k + (1-v_k)\frac{y_{k+1}}{y_{k+1}+\cdots+y_{n+1}},\ v_k + (1-v_k)\frac{y_{k+1}+y_{k+2}}{y_{k+1}+\cdots+y_{n+1}},\ \ldots,$$
$$v_k + (1-v_k)\frac{y_{k+1}+\cdots+y_n}{y_{k+1}+\cdots+y_{n+1}},$$

and it is obvious that, for given $v_k$, the random variables in (5) are distributed as the order statistics of $k-1$ uniform variables on the interval $[0,v_k]$, and are independent of the variables in (6), which are distributed as the order statistics of $(n-k)$ uniform variables on the interval $[v_k,1]$.

One may prove directly that each of $v_1,\ldots,v_n$ has a beta distribution, but an interesting alternative method is to take advantage of known facts concerning the ratios of quadratic forms in normal variables, as each of the $v$'s may be viewed as such. The relationship between the beta and the $F$ distributions may be brought in at this point, too.